NLP – Prediction Task

1. Twitter sentiment analysis

Commonly, people use social media to express their feelings and ideas. These ideas can contain sentiments or opinions which will be useful for a particular company, person or brand. If we consider Twitter, due to the huge volume of tweet generation, we cannot do this sentiment extraction manually. But we can easily do it using machine learning.

In this task, you are given a data set of tweets that are labelled as positive and negative. Using these labelled data, you need to train a machine learning model to predict the sentiment (positive or negative) of new tweets. You can use any machine learning algorithm you want to build the models, either from the sessions or your own research based on the topics we discussed.

Disclaimer: The text available with the data set can contain offensive language.

Note: Use 100 as the value for any random state/seed you set.

2. Data set

Within the data folder, there are 2 .csv files which are named “train.csv” and “test.csv”.

train.csv

Training data to use with model training and validation.
Number of entries: 16363
Columns:

- id – Unique Id
- text – Tweet text
- sentiment – Label assigned based on sentiment (positive or negative)

test.csv

Testing data to use with predictions.
Number of entries: 1000
Columns:

- id – Unique Id
- text – Tweet text

3. Submission

The final submission should include the following:

a. Test data predictions

Your test predictions need to be submitted to the **CodaLab**. Please carefully read the submission instructions given on the CodaLab page before submission.

Make sure you use the **best** model(s) you built to make predictions on the test dataset.

b. Blog explaining your implementation

A maximum 300-word blog explaining what you did and why including any plots and screenshots of the code needs to be submitted to **Moodle**.

Make sure to include evaluation results that you use to compare the models you built to select the best model in your blog.

Note: Both elements mentioned above **must be** submitted to consider your submission as a valid submission.

Hints:

Model evaluation

Use 70% of the training data to train the machine learning model and the remaining 30% as the validation set to evaluate the model.
The following code segment can be used to perform data splitting.

```
X_train, X_val, y_train, y_val = train_test_split(text_column, label_col
umn, test_size=0.3, random_state=100)
```

Save Predictions

The following code allows saving the predictions into a .csv file. Make sure to fill in the missing parts in this code according to your training process and use the correct model to make predictions.

```
# read csv file into a dataframe
df_test = pd.read_csv('/content/test.csv')

# Preprocess 'text' column in df_test using the same techniques you used
to preprocess training data
<add your code here>
# convert text to features using the same procedure used with training
data
X_test = <add your code here>

# make predictions using the best model
predictions = model.predict(X_test)

# add predictions to the 'prediction' column
df_test['prediction'] = predictions

# save data frame to .csv file
df_test.to_csv('/content/test-predictions.csv', index=False)
```

The saved .csv file can be used to generate the final submission file using the Python scripts provided with the CodaLab page.